Journal of Multivariate Analysis 152 (2016) 145–161

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

The use of a common location measure in the invariant coordinate selection and projection pursuit

Fatimah Alashwali [a,*], John T. Kent [b]

[a] Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia
[b] University of Leeds, Leeds, United Kingdom

ARTICLE INFO

Article history:
Received 25 March 2015
Available online 29 August 2016

AMS subject classifications:
primary 62G35
secondary 62H30

Keywords:
Cluster analysis
Invariant coordinate selection
Projection pursuit
Robust scatter matrices
Location measures
Multivariate mixture model

ABSTRACT

Invariant coordinate selection (ICS) and projection pursuit (PP) are two methods that can be used to detect clustering directions in multivariate data by optimizing criteria sensitive to non-normality. In particular, ICS finds clustering directions using a relative eigen-decomposition of two scatter matrices with different levels of robustness; PP is a one-dimensional variant of ICS. Each of the two scatter matrices includes an implicit or explicit choice of location. However, when different measures of location are used, ICS and PP can behave counter-intuitively. In this paper we explore this behavior in a variety of examples and propose a simple and natural solution: use the same measure of location for both scatter matrices.

1. Introduction

Consider a multivariate dataset, given as an $n \times p$ data matrix X, and suppose we want to explore the existence of any clusters. One way to detect clusters is by projecting the data onto a lower dimensional subspace for which the data are maximally non-normal. Hence, methods that are sensitive to non-normality can be used to detect clusters.

One set of methods based on this principle is invariant coordinate selection (ICS), introduced by Tyler et al. [17], together with a one-dimensional variant called projection pursuit (PP), introduced by Friedman and Tukey [5]. ICS involves the use of two scatter matrices, $S_1 = S_1(X)$ and $S_2 = S_2(X)$ with S_2 chosen to be more robust than S_1. An eigen-decomposition of $S_2^{-1}S_1$ is carried out. If the data can be partitioned into two clusters, then typically the eigenvector corresponding to the smallest eigenvalue is a good estimate of the clustering direction. The main choice for the user when carrying out ICS is the choice of the two scatter matrices.

However, in numerical experiments based on a simple mixture of two bivariate normal distributions, some strange behavior was noticed. In certain circumstances, ICS, and its variant PP, badly failed to pick out the right clustering direction. Eventually, it was discovered that the cause was the use of different location measures in the two scatter matrices. The purpose of this paper is to explore the reasons for this strange behavior in detail and to demonstrate the benefits of using common location measures.

Section 2 gives some examples of scatter matrices and reviews the use of ICS and PP as clustering methods. Section 3 sets out the multivariate normal mixture model with two useful standardizations of the coordinate system. Section 4

* Corresponding author.
E-mail addresses: fsalashwali@pnu.edu.sa (F. Alashwali), j.t.kent@leeds.ac.uk (J.T. Kent).

http://dx.doi.org/10.1016/j.jmva.2016.08.007
0047-259X/

demonstrates in the population setting an ideal situation where ICS and PP work as expected and where an analytic solution is available — the two-group normal mixture model where the two scatter matrices are given by the covariance matrix and a kurtosis-based matrix. Some examples with other robust estimators are given in Sections 5 and 6, which show how ICS and PP can go wrong when different location measures are used and how the problem is fixed by using a common location measure. Further issues, including unbalanced mixtures and heteroscedasticity, are discussed in Section 7.

Notation. Univariate random variables, and their realizations, are denoted by lowercase letters, x, say. Multivariate random vectors, and their realizations, are denoted by lowercase bold letters, $\boldsymbol{x}$, say. A capital letter, X, say is used for $n \times p$ data matrix containing p variables or measurements on n observations; X can be written in terms of its rows as

$$X = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top)^\top,$$

with ith row $\boldsymbol{x}_i^\top = (x_{i1}, \ldots, x_{ip}),\ i = 1, \ldots, n$.

2. Background

2.1. Scatter matrices

A scatter matrix $S(X)$, as a function of an $n \times p$ data matrix X, is a $p \times p$ affine equivariant positive definite matrix. Following Tyler et al. [17], it is convenient to classify scatter matrices into three classes depending on their robustness.

(1) Class I: is the class of non-robust scatter matrices with zero breakdown point and unbounded influence function. Examples include the covariance matrix defined below in (1) and the kurtosis-based matrix in (2).
(2) Class II: is the class of scatter matrices that are locally robust, in the sense that they have bounded influence function and positive breakdown points not greater than $1/(p+1)$. An example from this class is the class of multivariate M-estimators, such as the M-estimate for the t-distribution, e.g., [4,8].
(3) Class III: is the class of scatter matrices with high breakdown points such as the Stahel–Donoho estimate, the minimum volume ellipsoid (mve) and the constrained M-estimates, e.g., [7,18].

Each scatter matrix has an implicit location measure. Let us look at the main examples in more detail, and note what happens in $p = 1$ dimension. The labels in parentheses are used as part of the notation later in the paper.

The sample covariance matrix (var) is defined by

$$S = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top, \tag{1}$$

where for convenience here a divisor of $1/n$ is used, and where $\bar{\boldsymbol{x}}$ is the sample mean vector. The implicit measure of location is just the sample mean.

The kurtosis-based matrix (kmat) is defined by

$$K = \frac{1}{n}\sum_{i=1}^{n}\{(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top S^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top. \tag{2}$$

Note that outlying observations are given higher weight than for the covariance matrix, so that K is less robust than S. Again the implicit measure of location is just the sample mean. When $p = 1$, the scatter matrix $S^{-1}K$ reduces to 3 plus the usual univariate kurtosis.

The M-estimator of scatter based on the multivariate t_ν-distribution for fixed ν is the maximum likelihood estimate obtained by maximizing the likelihood jointly over scatter matrix Σ and location vector $\boldsymbol{\mu}$. If both parameters are unknown and $\nu \geq 1$, then under mild conditions on the data, the mle of $(\boldsymbol{\mu}, \Sigma)$, is the unique stationary point of the likelihood. Similarly, if $\nu \geq 0$ and $\boldsymbol{\mu}$ is known, the mle of Σ is the unique stationary point of the likelihood; see Kent et al. [8]. In either case, an iterative numerical algorithm is needed. Note that when $\boldsymbol{\mu}$ is to be estimated as well as Σ, the mle of μ is the implicit measure of location for this scatter matrix. For this paper we limit attention to the choice $\nu = 2$ (and label it below by t2).

The minimum volume ellipsoid (mve) estimate of scatter S_{mve}, introduced by Rousseeuw [14], is the ellipsoid that has the minimum volume among all ellipsoids containing at least half of observations, and its implicit estimate of location, $\bar{\boldsymbol{x}}_{\text{mve}}$, say, is the center of that ellipsoid. Calculating the exact mve requires extensive computation. In practice, it is calculated approximately by considering only a subset of all subsamples that contain 50% of the observations, e.g., [9,18]. If the location vector is specified, the search is limited to ellipsoids centered at this location measure.

When $p = 1$, the mve reduces to the lshorth, defined as the length of the shortest interval that contains at least half of observations. The corresponding estimate of location, $\bar{\boldsymbol{x}}_{\text{lshorth}}$, say, is the midpoint of this interval. Calculating the lshorth around a known measure of location is trivial; just find the length of the interval that contains half of observations centered at this location measure. The lshorth was introduced by Grubel [6], building on an earlier suggestion of Andrews et al. [3] to use $\bar{\boldsymbol{x}}_{\text{lshorth}}$, which they called the shorth, as a location measure.

The minimum covariance determinant estimate of scatter (mcd), S_{mcd}, say, is defined as the covariance matrix of half of observations with the smallest determinant. The mcd location measure, $\bar{\boldsymbol{x}}_{\text{mcd}}$, say, is the sample mean of those observations.